Panoptic Dynamic Voltage Scaling for Low Power Design

Yousef Shakhsheer, yas5b@virginia.edu

***Abstract*— Panoptic Dynamic Voltage Scaling (PDVS) extends DVS to finer granularity in space and time, allowing for much more flexible and energy efficient designs. Fine granularity results in higher energy efficiency by allowing non-critical path components to work at the lowest voltage possible while still meeting performance requirements. PDVS's headers can be divided and utilized as low drop-regulators without the expense of DC-DC regulators, thus providing more voltages for DVS.**

I. Introduction

Reducing energy consumption in CMOS circuit architectures continues to be an important area of research. Portable systems, such as cellular phones, medical sensor, and environmental sensors, demand extended battery lifetimes and high performance. There is a fundamental trade-off between power and speed in circuits. Designers can take advantage of the variable and bursty nature of these applications to potentially achieve both low energy and sufficient performance (completion of tasks by its deadline) rather than designing the system in a static fashion (running at high voltage at all times) to always meet performance constraints.

In recent years, the use of multiple supply rails, such as Multi-VDD (MV_{DD}), has been introduced to enable systems with strict latency and throughput requirements to still reduce energy by assigning lower voltages to components performing non-critical operations. Power-gating idle components reduces leakage energy. Dynamic voltage scaling (DVS) has also become commonplace, enabling systems to adapt to dynamic workloads and battery availability by operating at the slowest rate while still meeting performance requirements. DVS adjusts the supply voltage (V_{DD}) to match a circuit's workload, providing quadratic energy savings at lower processing rates when timing slack exists. Each of these techniques provides benefits individually. However, combining these approaches will provide maximum energy efficiency.

Recent DVS implementations limit the spatial granularity (ability to assign each component to different voltages at any given time) with which V_{DD} can vary to the core level or above. Existing DVS techniques also limit the temporal granularity (ability to adjust a component's voltage quickly) by relying on DC-DC converters to adjust V_{DD}, taking tens to hundreds of µsecs for an output voltage transition [1]. The coarse granularity of conventional DVS means that blocks can only save energy for wide ranging in workload.

Panoptic DVS (PDVS) extends DVS to finer granularity in space and time, allowing for much more flexible and energy efficient designs. Fine granularity results in higher energy efficiency by allowing components not on the critical path to work at the lowest voltage possible while still meeting performance needs on the critical path. Fine spatial granularity allows for fast switching from low to high supply voltage due to the lower virtual rail capacitance of smaller voltage islands. This paper presents PDVS, benefits of the traditional PDVS implementation, and using headers to generate more voltages.

II. Panoptic Dynamic Voltage Scaling

PDVS, shown in Figure 1, is implemented by routing multiple supplies throughout the chip, and using header switches at the component level which set the local V_{DD} to one of the routed supplies. This way, as number of components wanting an independent supply increase, the only increase in area is the headers. This is small as compared to a linear increase in DC-DC converters.

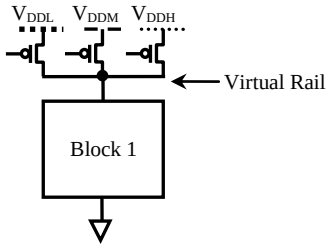

Fig. 1. PDVS architecture enabling local fine-grained DVS using header switches and a small set of shared V_{DD}s.

Header switches enhance traditional DVS capabilities by providing rapid and energy efficient transitions between processing voltages on clock edges with minimal energy overhead. PDVS also reduces area be reducing the number of components statically assigned to a voltage in schemes such as MV_{DD}. Instead, PDVS assigns voltages to operations.

PDVS provides two main benefits: fine spatial and temporal granularity. Like MV_{DD}, each component can thus use a voltage that allows it to achieve its required performance without consuming unnecessary additional power. The fine spatial granularity of this architecture gives static savings in energy whenever timing slack, a period in which no operation is taking place, exists since those components experiencing the slack can switch to lower voltages while still meeting their deadlines. PDVS provides this capability with lower area requirements than MV_{DD} by assigning voltages to operations rather than components. That is, the same component can execute at different voltages for different operations in the processing flow by switching to the appropriate power rail.

Header switches also enhance traditional DVS capabilities by providing rapid and energy efficient transitions between processing voltages on clock edges with minimal energy overhead. This includes the capability of rapidly switching from a high performance mode to an ultra low power mode (e.g. using sub-threshold operation) for periods of operation that have extremely relaxed performance constraints then back

to high performance mode again [2]. Measured chip results [3] show that the energy and time overhead of switching local voltages is small, allowing the system to take advantage of rapid changes in workload to save energy. However there are overheads associated with PDVS compared to single-V_{DD} and MV_{DD}. Adding headers and level converters, used to convert voltages between different blocks, results in a small area and delay penalties.

III. Benefits of PDVS

PDVS can improve the energy efficiency of a variable workload system even relative to an optimal conventional DVS scheme. Figure 2 shows how energy consumption scales with workload for different power management schemes [4]. For simplicity, workload is defined as the number of iterations of a loop that a system must run to complete a task in a given amount of time. The figure shows two sets of curves: global block and sub-block. Global block refers to the case where the entire system runs at the same V_{DD}. Within such a system there can be sub-blocks that have slack and thus could work at a lower V_{DD} with the entire system still meeting the workload requirements. Allowing sub-blocks to have separate V_{DD} connections defines the sub-block set of curves. With some blocks working at lower V_{DD}, the energy consumption of the system is lower without any system level performance penalty, giving the downward shift of the sub-block curves in Figure 2 relative to the global block curves.

Within both global block and sub-block schemes, there are three implementation possibilities: V_{DD} set by maximum workload (the linear curve), V_{DD} set by the current workload of the system (the dashed curve), and V_{DD} set as one of 2 possible choices, with the specific V_{DD} values being set by the designer (the piecewise linear curve connecting points that correspond to a specific V_{DD}). The first system must always operate at the maximum rate and enter a low-power idle/sleep mode (e.g. clock gate) when it finishes early. As Figure 2 shows, this provides only linear energy savings (i.e. a workload that is half of the maximum is executed with half of the maximum energy). On the other extreme, the second system can choose any V_{DD} based on the current workload, thus giving the ideal quadratic saving in energy. However, choosing any V_{DD} from a continuous range is not practical due to limitations in voltage converters. Even in converters with numerous voltage levels, the voltage transition latency is tens to hundreds of µseconds, thus preventing a system from saving energy in cases where workloads change rapidly. The third system uses two quantized values for V_{DD} and is also known as voltage dithering. Voltage dithering uses a small number of discrete voltages to approximate the ideal curves by operating for part of the time at a higher voltage and the remainder of the time at a lower voltage to average the performance at those endpoints and achieve any effective performance rate in between. Dithering thus leads to a linear characteristic that connects the dots between points on the ideal curve provided by the V_{DD} rails. Note again that these three sub-types can be realized in both global block and sub-block types of systems.

PDVS enables different sub-blocks of a system to work at different (amongst a few chosen) V_{DD} values. Thus PDVS is classified as a sub-block scheme. Further, because the header controls can be dynamically changed by microcode, PDVS enables voltage dithering. The above flexibility allows PDVS to achieve the lowest practical energy curve in Figure 2. However if need be, by making header controls static and/or making all V_{DD} values the same and/or set by the maximum workload, PDVS can still enable any power management scheme shown in Figure 2.

To demonstrate the savings theorized in Figure 2, [4] designed and simulated several benchmarks using single-V_{DD} (global block, linear curve), MV_{DD} (sub-block, linear curve), and PDVS. The benchmarks are designed in the form of dataflow graphs (DFGs) using adders and multipliers. For single-V_{DD}, all the adders and the multipliers use the same V_{DD}. For MV_{DD}, different adders and multipliers are permanently tied to any one of three V_{DD} values. PDVS provides temporal flexibility by allowing any adder or multiplier, at a specific time, to be connected to any one of the three V_{DD} rails. The specific implementation of the DFGs, including how operands are transferred to and from various components, will be discussed in the next section.

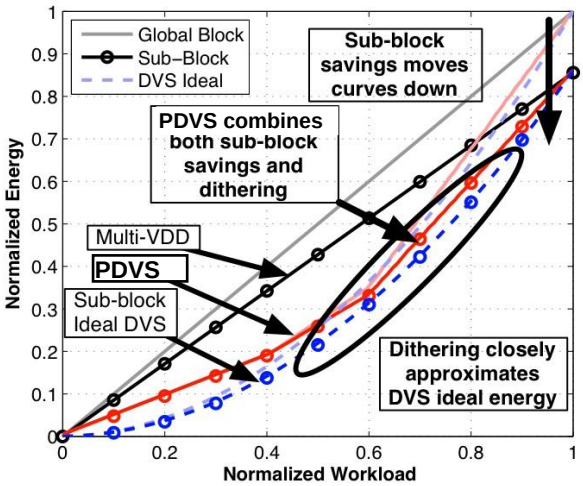

Fig. 2. From [4]. PDVS achieves sub-block energy savings and closely tracks ideal DVS with dithering.

A DFG can have adder or multiplier instances that are not on the critical path and have slack. These give an opportunity to save energy by using a lower V_{DD} setting. This corresponds to the downward shift of curves in Fig. 2. (sub-block energy savings) for any workload. Consider a system that only needs to execute a single DFG at a static rate. While the focus of such a system is not on dynamic workload changes, PDVS allows the use of the same component at three different V_{DD} values within the DFG. Thus given a latency constraint and either a fixed number of components or a maximum amount of energy, PDVS helps save energy when compared to MV_{DD}. PDVS can save area over MV_{DD} for a given energy since any component can be connected to any of the supply voltages. Therefore, a DFG that might require two adders at the high voltage and two at the low voltage in MV_{DD} can be run with two or three components in PDVS depending on the schedule.

IV. Other Applications of the PDVS Architecture

Circuits that utilize DC-DC converters for DVS incur significant delay and energy overhead for each voltage

transition. In addition, DC-DC converters make it impractical to have a large number of independently controlled DVS blocks, where each block can operate at its own locally-controlled voltage. Internal regulators require significant area, and external regulators require too many V_{DD} pins. As a result, most DC-DC converter-based DVS schemes employ global DVS in which the entire circuit operates at the same voltage.

Designers can utilize the existing PDVS architecture and split headers into smaller, parallel headers with individual gate controls (variable weighted headers). Variable weighted headers can be utilized as a simple, low dropout DC voltage regulator to efficiently provide local voltage control without adding new DC-DC regulators, changing the voltage output of the existing regulators, or adding metal routing. These headers provide the regulated voltage by utilizing the voltage drop across the header. There is no feedback loop to adjust the output voltage. Rather, the voltage rail settles during operation, thus providing the expected output voltage on V_{Rail}.

As header size decreases, the effective header resistance increases, thus reducing current through the header and causing the virtual voltage rail to droop to a lower value. This virtual rail voltage not only depends on the size of the header, but also on the activity factor of the block, data inputs of the block, and the full V_{DD} of the block. This virtual rail voltage sets the delay and energy of the operation.

Assuming frequent operation and no idling, the virtual rail will droop to a voltage below the nominal V_{DD}. Figure 3 shows the virtual voltage rail of successive operations of a 32b Kogge Stone adder for different percentages of the total header width. Notice that the rail settles near a certain voltage for each header size. In this operation, the energy is reduced by up to 37% in simulation as compared to circuit operation tied directly to V_{DD} with no header.

The energy of this operation using this approach is bounded by the following two equations:

(1)

(2)

with (1) and (2) providing the upper and lower bounds, respectively.

To highlight the potential benefits of using this droop to save energy, a ring oscillator was connected to variable weighted headers and allowed the rail to settle for each header width. Figure 4 shows the normalized energy and delay as the percentage of headers turned on varies, where 100% corresponds to all headers on (i.e. maximum header width). As expected, lowering the percent of headers on decreases the energy and increases the delay.

Additionally, as with DVS, there is an overhead associated with this variable weighted header technique. The virtual rail must eventually recover from the droop during operations. This recovery energy can be defined as

(3)

The system can amortize this recovery energy over multiple cycles by maximizing the number of operations run with this voltage droop, thus reducing energy per operation when the virtual rail is relatively constant. The gate capacitance also increases due to the separate gate controls and increase in the number of control wires required, thus increase gate switching energy.

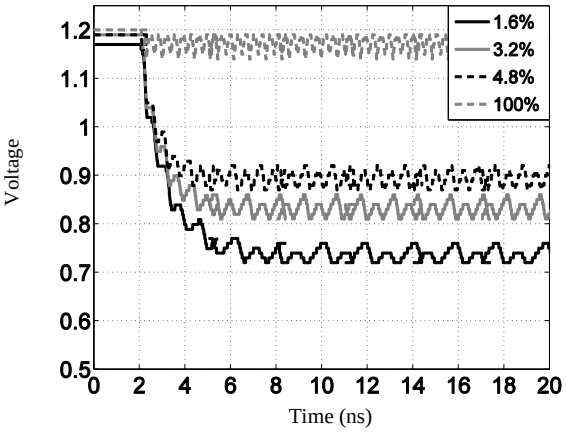

Fig. 4. Virtual rail of a 32b Kogge Stone adder during successive adder operations over several header widths.

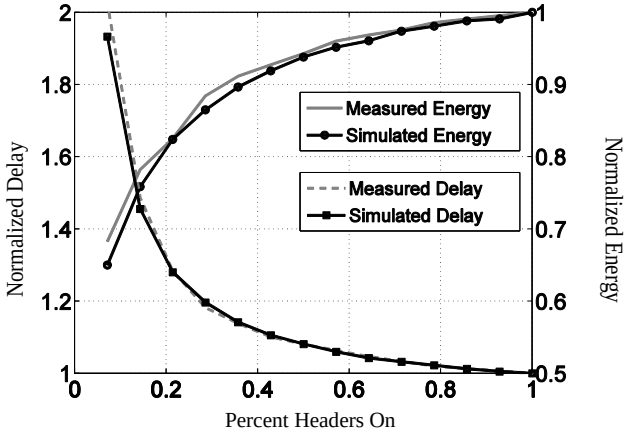

Fig. 4. Simulated and measured energy from 90 nm test chip and delay for a ring oscillator with sweeping header size.

V. Conclusions

PDVS, due to its fine spatial and temporal granularity, provides energy savings over architectures such as Single-V_{DD} and MV_{DD} with minimal overheads. Additionally, it provides area savings over MV_{DD}. The header in this architecture can also be modified slightly to be utilized as a low-dropout voltage regulator without the area, energy, and delay overheads of a DC-DC converter. These additional voltages can allow for more energy savings using the PDVS system.

References

[1] Zheng, C et al. "A 10MHz 92.1%-efficiency green-mode automatic reconfigurable switching converter with adaptively compensated single-bound hysteresis control." *ISSCC*, pp.204-205, 7-11 Feb. 2010

[2] Calhoun, B. et al. "Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering." *JSSC*, Vol. 41, No. 1, pp. 238–245, 2006.

[3] Di, L.et al. "Power switch characterization for fine-grained dynamic voltage scaling." *ICCD*, pp. 605-611, 2008.

[4] Calhoun,B. et al. "Flexible circuits and architectures for ultralow power." *Proceedings of the IEEE*, Vol. 98, No. 2, Feb. 2010.

[5] Truong, D. et al., "A 167-processor computational platform in 65 nm CMOS," *JSSC*, vol. 44, no. 4, pp. 1130-1144, April 2009